The GIAPSI NIST 2012 Speaker Recognition Evaluation System

L. M. Mazaira Fernández, A. Álvarez Marquina, P. Gómez Vilda

GRUPO DE INFORMÁTICA APLICADA AL PROCESADO DE SEÑAL E IMAGEN (GIAPSI)
Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, E-28660 Boadilla del Monte, Madrid - SPAIN
TEL: +34913367386; E-MAIL: luismiguel.mazaira@upm.es

INTRODUCTION

The system developed by the GIAPSI research group for the 2012 NIST Speaker Recognition Evaluation (NIST SRE12) takes as its starting point the system build for the NIST SRE2010, regarding the parameterization technique, but with a better fit thanks to the experience gained in the previous evaluation. Specifically, spectral characteristics of vocal tract (acoustic-phonetic) and glottal estimate (phonation-gesture) of voicing speech are combined with classical parameterization approaches based on the power spectral density of speech as a whole, to provide a compact representation for each speaker involved in the recognition process. Preliminary works [1][2], had shown that both vocal tract and glottal estimates bear essential biometric information, which can be applied in speaker recognition tasks. The GIAPSI 2012 system is no longer based on the classical GMM-UBM approach but in the GSV (Gaussian SuperVectors) and i-vectors paradigm.

FEATURE TEMPLATE

Voice Production

The physiological speech production model (Figure 1) assumes that voiced speech is generated by a glottal excitation signal *e(n)* which is spectrally conditioned by the vocal tract with transfer function given by $F_{vt}(z)$ to produce the speech signal before radiation $s_l(n)$ and after radiation *s(n)*.

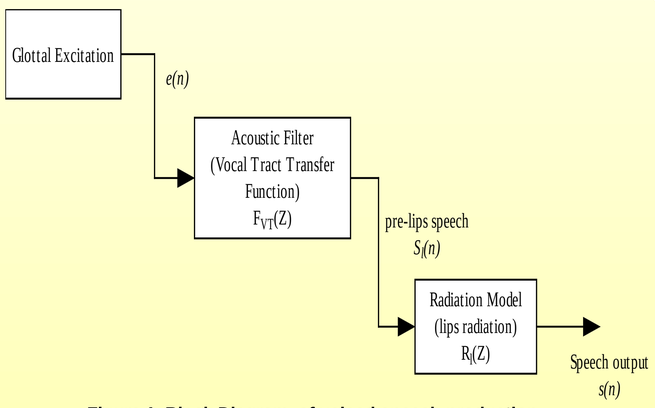

Figure 1. Block Diagram of voiced speech production

Source Separation

To estimate the glottal source from the speech signal, an iterative algorithm have been proposed. The block diagram of the applied method is depicted in figure 2.

The iterative algorithm comprises the following steps:

1. Remove the radiation effects from voice s(n), by filtering with $R_l^{-1}(z)$.
2. A k-order filtering process is applied to remove the vocal tract information from the radiation compensated speech ($e_g(n)$). This process can be implemented using a k-order prediction error lattice
3. The residual is used as the reference signal in an Adaptive Lattice-Ladder filter used for Joint-Process Estimation on the radiation-compensated speech, $s_l(n)$. Through this process, a glottal estimate, $s_g(n)$ and a vocal tract estimate, $s_v(n)$, are extracted which can be considered fully uncorrelated (second-order decoupling) [3]

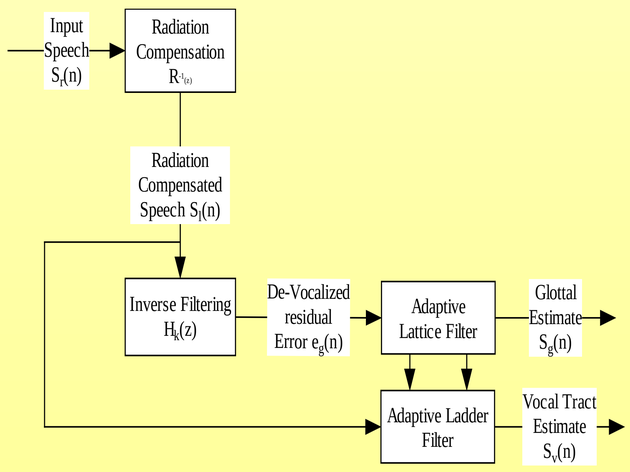

Figure 2. General framework for glottal source separation from voice by adaptive joint estimation

Feature Template

The most used feature sets for speaker recognition are the mel-frequency cepstral coefficients (MFCC) which are based on the magnitude spectrum of the speech. In the present system, MFCC are also used in the following configuration.

For voiced segments within each utterance, the glottal component is estimated and integrated. Using the speech signal and the glottal signal, a 45 feature vector has been extracted for each 32 ms voiced frame (with 8ms overlapping), which contains the following information which is gender dependent:

- MALE:
  - 20 MFCC + ΔMFCC
  - 2 MFCC + ΔMFCC derived from the vocal estimate
  - 4 MFCC derived from the glottal estimate
- FEMALE:
  - 18 MFCC + ΔMFCC (female)
  - 2 MFCC + ΔMFCC derived from the vocal estimate
  - 10 MFCC derived from the glottal estimate

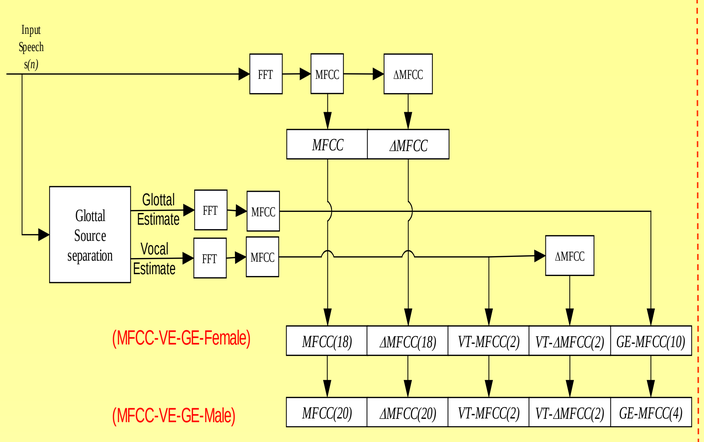

Figure 3. Parameterization scheme used in the GIAPSI System

SYSTEM DESCRIPTION

Preprocessing & Feature Extraction

A two step preprocessing stage is performed including a Voice Activity Detection stage (VAD) and a simple cross-channel speaker cancelation. An adaptive VAD algorithm based on energy detection has been implemented and computed over a 64ms-long Blackman window with 13ms overlap. In the case of speaker cancelation, the algorithm applied consists on removing segments on the channel of interest that match with segments classified as including voice in the complementary channel when 2 channels are available. Figure 4 depicts an example of this preprocessing stage (green area will be removed before template generation).

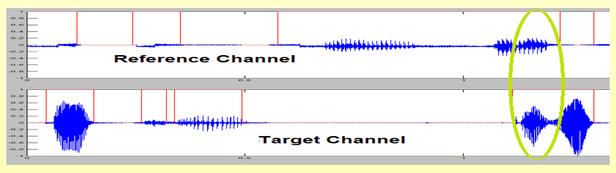

Figure 4. Cross-channel speaker cancellation example

Additionally, as some of the files include telephone conversations, it was necessary to perform a noise reduction preprocessing step. In this case, a variation of the Ephraim-Malah spectral subtraction algorithm in a single channel is applied. Channel distortion reduction techniques, such as Cepstral Mean Subtraction (CMS), Feature Warping (FW) and RASTA, has also been applied.

Speaker Modeling and Scoring

A block diagram of the Automatic Speaker Verification system implemented using the SV-GMM approach is shown in Figure 5. A gender-dependent UBM has been build (part *A* in Figure 5) via EM-algorithm using a specific training set. From this UBM, each speaker model has been adapted (part *B* in Figure 5), using the MAP algorithm in which only the distribution means have been adapted. The number of Gaussians as well as the relevance factor used on the MAP-algorithm depends on the specific experiment carried out (typically been 1024 and 18 respectively).
A vector to supervector mapping is then applied to transform the GMM model obtained by MAP adaptation into a supervector. In order to make the problem computationally efficient, a dimensionality reduction step has been added based on PCA. For the sake of simplicity, the PCA matrix training process as well as the data used to train this matrix has been removed in Figure 5.

Additionally, interssesion variability compensation must be applied. In our case, WCCN and LDA are combined.

As the test recordings can be transformed also into a supervector by the previously explained process (part *C* in Figure 5), a distance between two vectors must be defined in order to make a decision on whether the test recording has been uttered by the claimed speaker (part D in Figure 5). The Cosine distance is used in the decision step. To make speaker recognition task more robust the ZNorm and TNorm score normalizations have been applied.

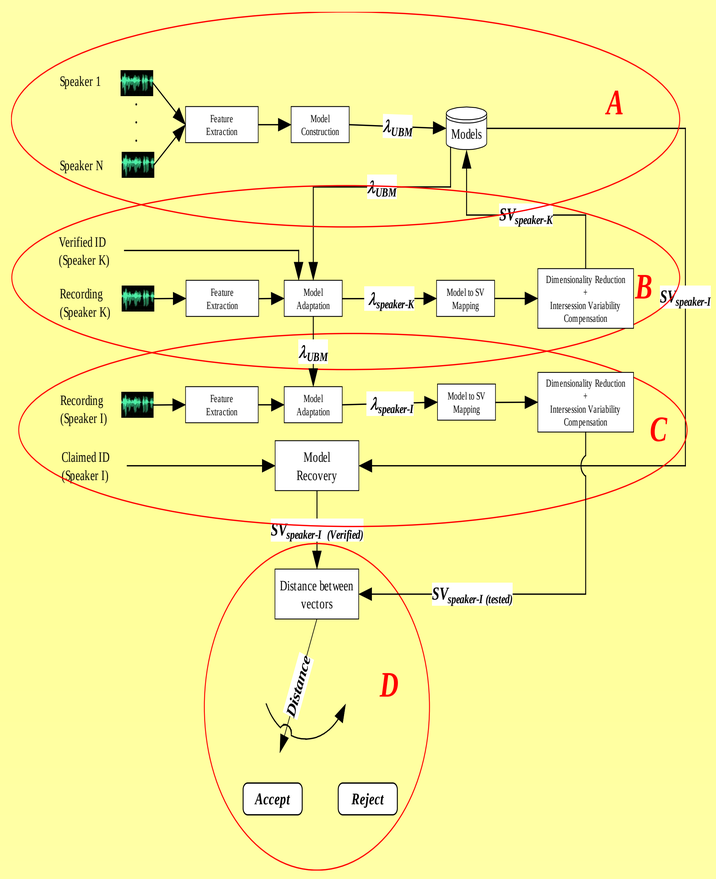

Figure 5. Block diagram of NIST SRE2012 GIAPSI System

CORE EVALUATION RESULTS

Results on core-core condition

The following set of figures shows the results achieved for the different conditions in which the core-core trial is divided. Specifically, we compared the results achieved both for the presented system based on the GSV paradigm and for the alternative developed system based on the i-vector paradigm.

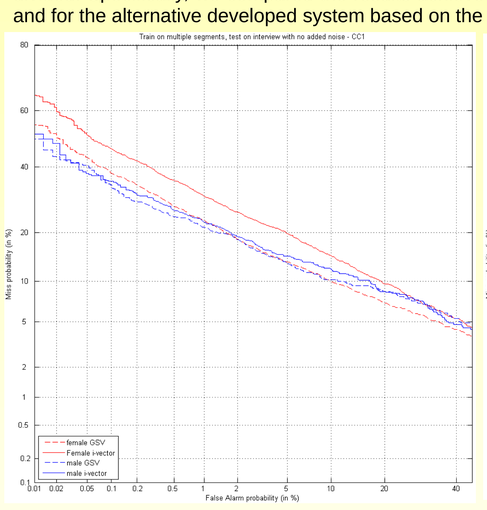

Figure 6 DET curve condition 1

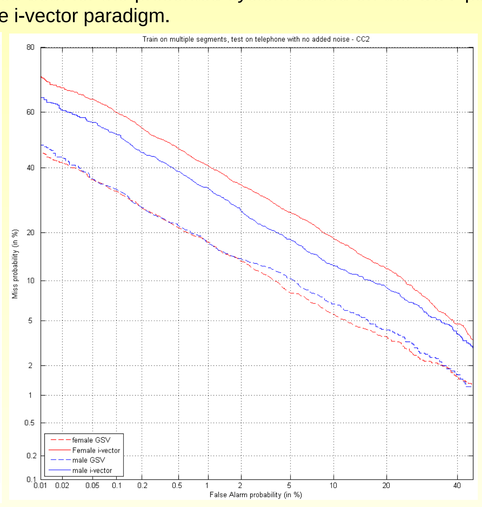

Figure 7 DET curve condition 2

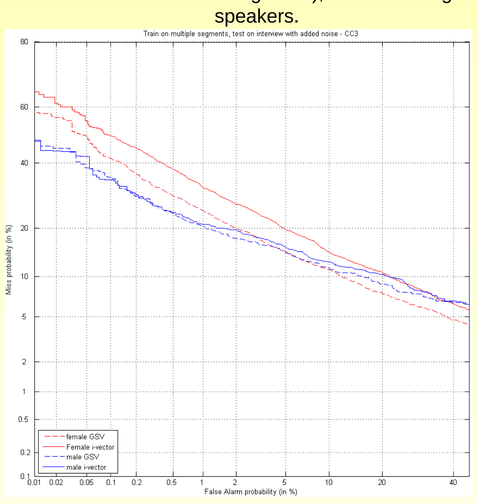

Figure 8 DET curve condition 3

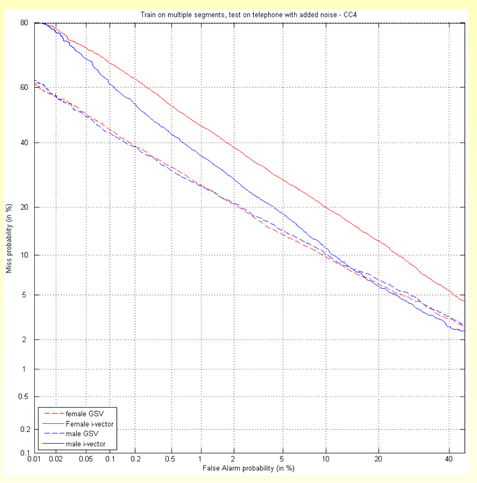

Figure 9 DET curve condition 4

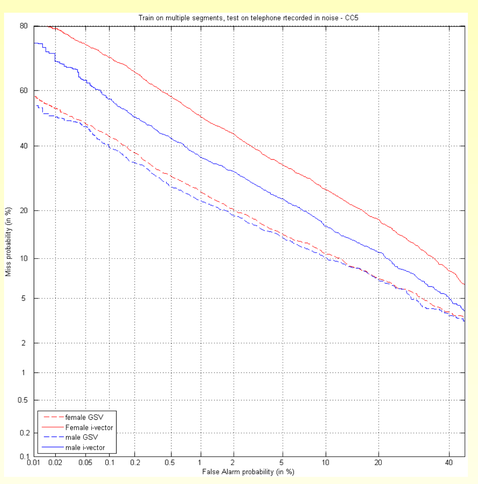

Figure 10 DET curve 5

Gender influence

The use of a gender-dependent biometric parameterization allows for a similar recognition rates for both genders (specially in the case on using GSV), which is of great interest considering that usually SR systems tend to have more trouble identifying female than male speakers.

CONCLUSIONS

NIST SRE provides a suitable framework to test the feasibility of using alternative parameters, different from the classic MFCC features, for characterizing a speaker in somehow real life scenarios. In particular classical parameters have been complemented with information obtained from spectral parameterization of the vocal tract and glottal estimates. Although the proposed algorithm do not provide a glottal estimate that strictly follows the well-known Liljencrants-Fant model, it is clear that the biometric information recovered from it as well as from the vocal tract estimate helps to improve the speaker recognition rate, thus confirming the conclusions in previous works [4]. Additionally, the results achieved in the NIST SRE2012 (EER<10%) shows a great improvement in terms of EER if compare to the results achieved by the group in the previous NIST SRE2010 (EER≈30%)

However, the results obtained in the NIST SRE2012 are still far away from the best results achieved by other researchers. This may be due to several reasons:

- No side information from recordings has been used, neither during training nor during testing.
- More preprocessing work is needed in the reconstruction of the vocal and glottal estimates, as the systems must deal with some of these factors: extremely noisy recordings, low-quality telephone recordings, very low-recording levels, etc.
- Limited amount of complementary databases (speakers and channel variability) to perform normalization. This is specially important in the case of using the i-vector approach.

Future work will also include the analysis of different classification methods as well as a post-processing of the data in order to compare these results with the ones obtained in the case of not using biometric parameters.

REFERENCES

[1] Gomez, P. et al., “A hybrid parameterization technique for speaker identification”, *In 16th European Signal Processing Conference* (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008.
[2] Gómez, P., et al., “Biometrical Speaker Description from Vocal Cord Parameterization”, *Proc. Of ICASSP’06*, Toulouse, France, 2006, pp. 1036-1039.
[3] Haykin, S., Adaptive Filter Theory, (4th Ed.), Prentice-Hall, Upper Saddle River, NJ,2001
[4] Mazaira, L.M, et al, “Improving Speaker Recognition Rates using alternative gender-dependent MFC coefficients”, Proceedings of the VI Jornadas de Reconocimiento Biométrico de Personas JRBP12, Spain 2012, pp 207-216.